9

COMMON EXAMINEE METHODS

Robert Coe

Abstract

Aim

The chapter aims to provide a full description of the methods that have been used to monitor comparability between different examinations taken concurrently by the same candidate: common examinee methods. Criticisms that have been made of these methods are presented and discussed in relation to different interpretations and conceptions of comparability.

Definition of comparability

Three conceptions of comparability are considered: *performance*, in which comparability is defined in terms of observed phenomena in relation to specific criteria; *statistical*, in which comparability depends on an estimate of the chances of a particular grade being achieved; and *construct*, in which examinations are seen as indicating different levels of a common linking construct.

Comparability methods

A number of methods are described, including simple comparisons of the grades achieved in a pair of examinations (subject pairs analysis); the aggregation of such paired comparisons to compare a larger set of examinations; the Nuttall *et al*. (1974) 'unbiased mean total' method; Kelly's (1976a) method; analysis of variance; average marks scaling; Rasch modelling.

History of use

Most of these methods have been used since the 1970s, though their prominence appears to be less now than in those early days, at least in England.

Strengths and weaknesses

Some of these methods require specific assumptions, such as equal intervals between grades. Other assumptions (for example, that examinations must be uni-dimensional, or that differences in factors such as teaching quality or motivation must be ignored) may depend on how the methods are applied and interpreted. Hence, evaluating the strengths and weaknesses of the methods is complex. Most of the methods considered, however, are capable of valid interpretation under the right conditions.

Conclusion

Common examinee methods should be part of the systematic monitoring of standards across syllabuses, boards and subjects, though, like all comparability methods, they should be used with caution and judgement. However, their main value is probably in informing the interpretation and use of examination results for purposes such as selection.

1 Introduction

Common examinee methods – comparing the achievements of the same examinees in different examinations – have been at the heart of debates about comparability of standards for many years. Some of these methods are quite simple to do and their results may also seem seductively simple to interpret. If common candidates typically and routinely get higher grades in one subject than they do in another, surely this means that the latter is harder?

On further reflection, however, such simplistic interpretations are both unclear and unjustified, at least in a strict sense. A number of writers on this subject have provided strong critiques of such over simplistic interpretations. From these it is clear that there may be many reasons why candidates achieve different grades in different subjects, apart from the obvious one that some are harder. It is also apparent that defining precisely what we mean by 'harder' in this context is very problematic.

Nevertheless, the fact appears to remain that the grades achieved in different subjects by the same candidates are in some cases robustly and substantively different. Unless we can find some convincing explanation for this phenomenon we are left with what appears to be an anomalous and even undesirable situation.

In this chapter I will first outline the general approach of common examinee methods with some specific examples. The different methods that have been used are then presented in detail, followed by an outline of the general criticisms that have been made of them. In trying to evaluate these criticisms it is then necessary to consider a number of different ways in which the concept of 'comparability' might be interpreted, and hence the different ways in which words like 'difficulty' or 'standards' might be understood. I also try to explore their logic and assumptions in relation to specific uses that have been made of their results. A full defence of the methods against these criticisms is then presented.

In trying to structure this chapter it has been impossible to find a way of presenting all the key ideas in logical sequence. In order to discuss any of the strengths and weaknesses of common examinee methods one must first have an understanding of what those methods are. This requires some understanding of the kinds of assumptions that must be made to use them, and hence of their appropriateness. However, one cannot really discuss the critiques of these methods without first making some conceptual clarifications, the justifications for which require an appreciation of some of the confusions that have arisen in using, interpreting and criticising the methods which is hard to do before the methods have been presented in detail. The reader may also find it helpful to refer to the commentary on Chapter 4, in which the different conceptions of 'comparability' are discussed. It is hoped that the resulting structure of the chapter is not too confusing or incoherent.

2 Background to common examinee methods

2.1 Origins of these methods

The use by qualification awarding bodies of statistical comparisons to inform the process of setting and monitoring grade standards has a long history, and common examinee methods were among the first to be used. Certainly as far back as the early 1970s, subject pairs analysis (SPA) was being widely used (Nuttall *et al*., 1974). Indeed, such methods are still used today, both to set grade standards and to monitor comparability, though the emphasis placed on statistical comparisons across subjects in setting thresholds for the award of particular grades in England, Wales and Northern Ireland may be less (Jones, 2003).

This decline may be traced to the growing emphasis through the 1980s on criterion referencing and the belief that standards could be specified and described in absolute terms without having to depend on comparing performance to statistical norms. Criticisms of common examinee methods began to appear with their earliest uses (Bardell *et al*., 1978), but a flurry of criticisms of the methods and assumptions of SPA (e.g. Willmott, 1995; Alton and Pearson, 1996; Goldstein and Cresswell, 1996; Newton, 1997) emerged in England in the mid-1990s, perhaps in response to some high profile uses of the method (Fitz-Gibbon & Vincent, 1994; Dearing, 1996).

Many of these critical authors were associated with the examining boards, though some were also users of the method (e.g. Sparkes, 2000). By the end of the 1990s a consensus appeared to emerge among the awarding bodies in England, Wales and Northern Ireland that the problems of these statistical methods for establishing comparability were so great that, as in the subtitle of Newton's (1997) paper, 'statistical techniques do not make the grade'. Instead, the use of examiners' judgements was held to provide a basis for ensuring comparability (Cresswell, 1996).

Although common examinee methods have most often been used in comparing different subjects, they have by no means been limited to this. They have, for example, also been applied to comparisons of different syllabuses in the same subject, or of different examining boards or even of different modules or units within a syllabus (e.g. Bardell *et al*., 1978; Backhouse, 1978; Forrest & Vickerman, 1982). Nevertheless, most features of the methods can be well illustrated by considering the case of subject comparisons and for the sake of simplicity most of the discussion in this chapter will refer to that case.

2.2 Outline of the general common examinee approach

All the methods described in this chapter broadly share the same underlying logic. They all set out to provide some estimate of the relative 'difficulty' of different subjects or examinations taken at the same time. They all compare the achievement of a given candidate in one examination with that same candidate's performance in one or more other examinations, and then to aggregate such comparisons across all available candidates. All these methods appear to rest on the assumption that if all examinations were equally difficult, other things being equal, one would expect that the grades a candidate is likely to achieve should not depend on which particular examinations they take. Implicit in this must be some rationale for comparing different subjects, though such a rationale often seems implicit rather than explicit. In order to understand what is meant by comparability, we must explicate this notion in rather more detail.

An example: subject pairs analysis

Probably the most widely used common examinee method is SPA. Indeed a number of critics of the general approach have taken this method as a representative example of the whole set of methods. In fact there are a number of different versions of SPA (see below) but it may help to describe the logic of one version to illustrate the general method.

If we pick two subjects to compare we can consider all candidates who have taken both. We then simply calculate the difference between the mean grade achieved by those same candidates in each subject. If they typically achieve better grades in one than the other we may say that the former is 'easier', the latter 'harder' – though precisely what these words mean in this context will be discussed at some length below.

A specific example[1] will illustrate the logic of this method, using data from the National Pupil Database for England. In 2004, 583,300 candidates entered both English and mathematics GCSE examinations. Table 1 shows the percentage of this group that achieved each of the available grades. If we assign each grade the value shown in the table (making the assumption that the intervals between grades are equal, both within and between subjects), we can see that for these candidates the average grade achieved in English was 4.76, compared with 4.46 in mathematics. In other words, the same candidates typically achieved 0.3 of a grade better in their English examination than in their mathematics. As these were the same people in both examinations, we know that characteristics such as their prior attainment, gender, social background and the type of school they attended are matched perfectly, so presumably none of these factors can account for the difference.

Table 1 Percentage achieving each grade

Grade	Value	Percentage achieving that grade in English	Mathematics
U	0	1.7	3.0
G	1	2.4	3.5
F	2	5.7	8.7
E	3	11.0	14.9
D	4	19.3	17.1
C	5	25.5	22.0
B	6	19.2	18.3
A	7	11.5	8.1
A*	8	3.7	4.4

One possible remaining explanation of this difference is that the mathematics examination was more ‘severely graded’ than the English, and it might be tempting to conclude that this seems the most obvious explanation. However, it might be that the quality of teaching experienced by those candidates in mathematics was not as good as in English; or they might simply have enjoyed English more and so worked harder. Either way, they would have genuinely learned and achieved more and so deserved their better grades. A number of other explanations are possible for this phenomenon and will be discussed below. Moreover, there are some significant conceptual problems with the meaning of terms like ‘harder’, and several alternative ways of calculating this difference. Hence it would be fair to say the phenomenon is rather more problematic than it first appears.

One specific problem is that even if we believe we can interpret the result as showing that mathematics is harder than English, we have really only shown that this is so for the group of candidates who sat both examinations. In the case of this particular subject pair the problem may not seem too great since 93% of those who entered any GCSE examination took both mathematics and English. For other pairs, however, it seems a good deal more serious. For example, both physics and media studies GCSEs had substantial entries (43,000 and 35,000 candidates respectively), but fewer than 1,300 candidates nationally entered both these subjects. Knowing that for this small proportion, physics was 0.3 of a grade harder than media studies does not seem to tell us much about the relative difficulties for all those who entered (or who might have chosen to enter) each subject.

A further problem is that the relative difficulty of the two subjects calculated in this way varies substantially for different sub-groups. For example, if the comparison between mathematics and English is limited to female candidates who took both subjects, the difference increases to 0.6 of a grade. For males, the difference is only 0.04 of a grade. In other words, for male candidates mathematics and English were about equally difficult, while for females mathematics was substantially harder. When both sexes are combined we get an averaged estimate of the difference in difficulty between the two subjects. It seems counterintuitive that the relative difficulty of two subjects should depend on the population of candidates who happened to take them, rather than being a feature of the examinations themselves. Different estimates of relative difficulty for different sub-groups present a challenge to our understanding of what we mean by ‘difficulty’, as well as forcing us to consider carefully the sample of candidates on which our estimates are based and to ask of what population it is (or should be) representative. We will return to a discussion of these issues below in the context of criticisms that have been made of the approach.

The appeal of the method

Despite these problems, this kind of approach has a strong appeal. By comparing performance in a particular subject with that same candidate’s performance in their other subjects, common examinee methods provide an obvious and superficially compelling control for differences in the general characteristics of individual candidates. Any factor that is a general feature of the individual candidate, such as their prior attainment, gender, social and ethnic background or 'general' ability is automatically matched perfectly. This also applies to any characteristics of the school which the candidate attended, such as its funding status (independent vs. state) or the social, gender and academic mix of its students. Of course, factors that are specific to that candidate in a particular examination, such as motivation or the quality of teaching experienced, are not controlled for.

Common examinee methods generally require no other data to be collected, and no matching with other datasets;[2] hence they may be possible when other methods are not. There is no need to worry about how reliable the measures of any covariates are, since they need not be directly measured. The lack of any reference test (or the use of other covariates) avoids the objection that results of comparisons may be sensitive to the particular kind of test used, or to arbitrary decisions about what variables to include in the model. And of course in practice which variables are included in such models is often not so much a matter of decision as of making the best of what is available.

The best way to illustrate the appeal of the method may be with an example of its application. Analysis conducted by the Advanced Level Information System (Alis, 2004) on A level results over a ten-year period shows that on average, students who take psychology A level achieve 0.9 grades higher in that subject than comparable students who take chemistry.[3] This is a sizeable difference, especially given its consistency over such a long period. One of the main uses of A level grades is as an entrance qualification to higher education. Suppose an admissions tutor in, say, history or economics has to choose between two otherwise equivalent candidates with the same grades, one of whom has taken chemistry, the other psychology. Should the differences in typical performance in these subjects be taken into account?

It seems reasonable to assume that an admissions tutor in an unrelated subject is not particularly interested in the specific skills accredited by these examinations, but interprets them as an indicator of a student's generalisable capacity for learning in other academic contexts. In this case, it might be true that psychology has been better taught than chemistry; candidates in the former might genuinely deserve their better grades. But if so, the higher grades may not reflect a greater capacity for future learning in a different context. Equally, students who chose psychology may have had a special talent or passion for it and so again deserve their higher grades, but again these qualities may not transfer to their future learning in a different context. We can carry on trying to think of reasons why students might have done so much better in one subject than the other, but in all cases the interpretation seems to be the same. Whatever the reasons for better performance, if we want to get a fair indication of a candidate's suitability for a non-overlapping course, those grades in psychology should be reduced by 0.9 to make them comparable with chemistry.

If the logic of this argument is accepted then there is a compelling case for interpreting the statistical differences from common examinee methods as indicating something about the relative value of grades achieved in different subjects. Whether we might want to say that chemistry is 'harder' than psychology is more problematic.

3 Detailed explanation of the different common examinee methods

All common examinee methods estimate some kind of statistical difference between the grades achieved in one subject and the grades achieved by the same candidates in one or more other subject(s). Comparisons may be limited to those achieving a particular grade in one of the subjects, or they may involve some kind of averaging process. There may also be an attempt to quantify the size of the difference. Within these broad similarities, however, there are a large number of different variations of the method.

3.1 Subject pairs analysis and its variants

Pair-wise comparisons of two subjects

The simplest version of all common examinee comparability methods is to consider only those candidates who have taken a particular pair of examinations. For each candidate we could determine whether they have achieved the same grade in each or, if not, in which subject they have done better. Simply counting the proportion of candidates in each category would form the basis of a comparison between the two subjects. An example of this approach can be found in an early report by the UK GCE Examination Boards (1971) which used the fact that in those days candidates might enter more than one board's examination in the same subject.

It may be, however, that, rather than making an overall comparison between subjects, we wish to focus on the performance of candidates who achieved a particular grade in one or other of them. For example, we could consider all candidates who achieved a grade C in Subject 1 and calculate the proportion who achieved the same grade or better in Subject 2, then compare this with the equivalent proportion when the subjects are reversed. This approach is described by Fearnley (1998) as the 'by grade' method used by the Northern Examinations and Assessment Board (NEAB), based on 'conditional cumulative percentages of candidates by grade' (method A, p. 13). This method therefore ignores all candidates who did not achieve a grade C in at least one of the two subjects in order to limit the comparison to this grade.

An intermediate approach, which focuses on a particular grade but slightly less narrowly (Fearnley's method C), is to consider all candidates entering both subjects and calculate the cumulative percentage achieving each grade in each subject. At each grade the percentages can be compared. This method (described by Fearnley, 1998, as 'marginal cumulative percentages of grades') allows a comparison to be made at a particular grade, but includes in the calculation the results of all candidates who achieved at least that grade in either subject.

The methods described so far depend on the ordinal property of grades, but not on any interval property. This, of course, is a strength, since assigning numbers to grades is essentially arbitrary and may misrepresent the relative sizes of the gaps between them (Fowles, 1998). However, a weakness of such ordinal approaches is that they are hard to aggregate. If we are prepared to make a stronger 'interval' assumption about examination grades, then we may be able to make comparisons that can more easily be combined.

One way to do this is to compute an average difference in the grades achieved in the two subjects. These methods may be described as ‘interval’ approaches since any such average will be sensitive to the sizes of the gaps between grades, not just to their order. The conventional way to do this is to convert examination grades into a numerical scale using consecutive integer values (e.g. at GCSE, U=0, G=1, F=2, …, B=6, A=7, A*=8). For each candidate who took a particular pair of subjects we can calculate the difference between their grade in Subject 1 and Subject 2. The mean of these differences across all candidates is a measure of the difference in ‘difficulty’ of the pair, in grade units. This method has been widely used to compute pair-wise comparisons (e.g. Forrest & Smith, 1972; Nuttall *et al.*, 1974, Ch. III). Forrest & Vickerman (1982, p. 9) note that changing the numerical scale to one that reflects the different mark intervals between grade boundaries produces results which are ‘almost indistinguishable from those obtained by the simple method’, suggesting that the interval assumption is not too problematic for these methods.

It is also possible to limit this average to those candidates who achieved a particular grade in each subject in order to compare the subjects specifically at that grade. This is Fearnley’s (1998) method B, ‘conditional mean grade by grade’.

A problem with all these variants of SPA, however, is that we can compare only two examinations at a time. One subject can be compared with a second, or indeed with any other in which it has candidates in common, but the result would be a set of pair-wise comparisons between subjects, rather than any kind of overall comparison of a larger set of subjects. Each of these pair-wise comparisons would be limited to the group of candidates who had taken both subjects; for most pairs, this group is unlikely to be representative of all those who take either subject. For example, a comparison between, say, Latin and media studies could be based only on those who had taken both subjects; other than by considering this small and atypical group we would not be able to say anything about the relative ‘difficulty’ of these two subjects.

A further limitation of such pair-wise comparisons is that some of the pairs of examinations we might want to compare have no candidates in common. For example, in England there is currently a common timetable so that all boards’ examinations in the same subject take place at the same time. Comparisons of syllabuses from different boards in the same subject could therefore not be based on candidates who had entered both. Hence the comparison must be more indirect: comparing each syllabus with the same group of other subjects may allow us to compare the two.

Aggregated subject pairs analysis

If we are prepared to adopt one of the ‘interval’ approaches it is relatively straightforward to calculate the mean grade differences for all possible pairs of subjects and to average the mean differences, for each subject separately. So, for example, we can calculate the average difference in the grades achieved in mathematics and every other subject taken with it. An average of these differences will give an estimate of the overall 'difficulty' of mathematics, compared with all the other subjects taken by candidates who also took mathematics. If we do this for all subjects, we arrive at a list of subjects with an estimate of relative 'difficulty' for each.

This approach is essentially method (iii), described by Alton & Pearson (1996). A full description of the method can also be found in Forrest & Vickerman (1982), along with detailed results from its application. A further example comes from Newbould & Schmidt (1983) who compared the grades achieved by all the candidates who took A level physics with the grades achieved by those same candidates in their other subjects. After doing this for two different syllabuses, physics and Nuffield physics, they concluded that 'the former was too difficult' (p. 20).

This method of averaging the pair-wise comparisons between subjects to get an overall comparison for all subjects is sufficiently different from the simple pair-wise SPA that it seems important to distinguish it. For this reason this approach will be denoted here as 'aggregated subject pairs analysis' (ASPA). Most reported uses of the aggregated approach, however, simply refer to it as subject pairs analysis, so there is plenty of scope for confusion here.

One key difference between SPA and ASPA relates to the problem of representativeness. A criticism that has been made of SPA (mentioned above, but see below for further discussion) is that the group of candidates who enter both subjects in a particular pair may not be representative of those who take either. In ASPA, however, the estimate of the relative 'difficulty' of a subject is based on the relative achievement of all candidates who took that subject with any other subject in the set being compared, so the problem of representativeness appears much less.

A refinement of the basic ASPA method is to use a weighted average in aggregating all the pair-wise differences for a particular subject. Clearly, some of the subject pairs may be expected to have many more candidates than others and it may seem inappropriate for combinations with very few candidates to make the same contribution to the estimate of a subject's difficulty as combinations that are far more popular. Hence each pair's difference is weighted by the number of pairs on which it is based. On the other hand, one could argue that since the analysis is really about subjects, not candidates, each subject should count the same in the calculation. A discussion of this issue can be found in Nuttall *et al*. (1974, p. 48).

An alternative weighting method is to weight each subject pair's difference by the correlation coefficient for the agreement between the grades achieved in the two subjects. In this approach, the estimate of a subject's 'difficulty' would depend more on comparisons with 'similar' subjects than on comparisons with relatively disparate subjects.

Subject triples and other combinations

A logical extension of the idea of comparing two subjects by considering all those who entered both is to do the same for three or more subjects. The analysis of subject triples takes three subjects and the set of candidates who have taken all three. Comparisons can then be made among the three, using any of the methods outlined above for subject pairs. This approach has been used, for example, by Alton & Pearson (1996). The analysis of triples is somewhat more complex than pairs, given the much larger number of possible combinations. Even with large data sets, some triples are likely to have quite small numbers, making them potentially unstable. And the extent to which the group of candidates who have taken a particular triple are representative of those who have taken each of the three subjects is also likely to be less than for pairs. Perhaps for these reasons, this approach does not seem to have been widely used.

Some of the uses of SPA have limited the groups of candidates to be considered in different ways. For example, Newbould & Schmidt (1983), in comparing physics with chemistry using SPA, split the analysis according to what other subjects were taken with the pair (e.g. comparing physics with chemistry for those also taking biology gave quite different results from the comparison for those also taking mathematics with the pair). Massey (1981) did a similar analysis for A level English, comparing English with other subjects for different third-subject combinations. Other studies have used this approach to compare the performance in physics of those who also take mathematics with the performance of those who do not (e.g. Rutter, 1994; Ireson, 1996).

3.2 Subject 'matrix' methods

A number of common examinee methods base their analysis on what may be thought of as a matrix of the results of candidates by examinations. These methods are all quite different, but may be grouped under this heading for convenience. It could also be argued that the ASPA method belongs with this group, though it has been presented above as a continuation of simple SPA.

Use of average performance as a reference

There are a number of different variations on this idea, which uses performance in a set of other subjects as a comparator for the grade achieved in a particular subject. Whereas ASPA estimates the 'difficulty' of a base subject by considering each other subject paired with it in turn, then averaging, by contrast, these approaches conduct some kind of aggregation first, then make the comparison.

One example is presented by the Welsh Joint Education Committee (WJEC, 1973). This analysis was limited to those candidates who had entered precisely eight of the eleven subjects compared. For each subject, a regression line was drawn to show the relationship between the grade in that subject and the grades achieved by that same candidate in all their other subjects. The average grade in the subject was also compared with the average achieved in all the other subjects. This is exactly equivalent to the use of SPA, differing only in the order of aggregation.

A second example is the Nuttall *et al*. (1974) 'UBMT method' (UBMT denotes 'unbiased mean total'). This method is genuinely different from SPA and, unlike the previous one, can be applied to candidates who have taken different numbers of subjects (provided they have taken more than one). As Nuttall *et al*. explain (p. 32): 'For all candidates taking a subject, e.g. chemistry, UBMT is the mean grade of all other subjects attempted for all candidates taking chemistry.' The method consists in first calculating the average grade achieved in chemistry by all candidates taking chemistry and at least one other subject. Then, for the set of all these candidates, the mean grade across all the other subjects they have taken is calculated. Finally, the averaged mean grade achieved in all the other subjects taken by these candidates (i.e. the UBMT) is subtracted from the average chemistry grade to give an index of the 'difficulty' of chemistry. The same process is repeated for the other subjects in the set. The equation for this model is given in the appendix to this chapter.

Kelly's method

One possible objection to the use of the UBMT method described above is that candidates who take a particular 'hard' subject may be likely to combine it with other 'hard' subjects; and, similarly, 'easy' subjects are more likely to be combined – and hence compared – with other 'easy' subjects. This could lead to the extent of the differences between subjects being underestimated.

For example, if a high proportion of those who took chemistry (a relatively 'hard' subject) also took other 'hard' subjects like mathematics and physics, the average grades they achieved in their other subjects might be quite similar to their grades in chemistry. Methods such as UBMT or ASPA (especially the weighted version) would then estimate chemistry to be of only average difficulty. Kelly's (1976a) method essentially uses an iterative procedure to respond to this problem.

The method begins in the same way as the UBMT method, by comparing the grades achieved by candidates in one subject with their average grades in all their other subjects, and so estimating the 'difficulty' of that subject. This is done for each subject under consideration, using the grades achieved by all candidates who have taken it with at least one other in the set. These 'difficulty estimates' are then used to apply a correction factor to the grades achieved in that subject. So, for example, if chemistry is found to be half a grade more 'difficult' than the average, that half grade is added to the achieved grade for all chemistry examinees. The whole process is then repeated using the 'difficulty corrected' grades in each subject instead of the actual achieved grades, to produce a new estimate of the relative 'difficulty' of these subjects with corrected grades. After a small number of iterations, the corrections shrink to zero and so the estimates of 'difficulty' of each subject converge.

Although it may be conceptually helpful to think of this method as iterative, it can be shown that the result is equivalent to solving a set of linear equations (Kelly, 1976a, provides a proof of this, due to Lawley, in an appendix). In practice, solving these equations using a matrix inversion is more efficient than the iterative process for large data sets. The full equations provided by Lawley are presented in the appendix to this chapter.

This method has been used relatively frequently, though perhaps by quite a limited number of researchers. The first reported use was by Kelly (1976a) in Scotland. It has subsequently been used by Fitz-Gibbon & Vincent (1994) and by other researchers at Durham University's Curriculum, Evaluation and Management (CEM) Centre (e.g. Alis, 2004; Yellis, 2006), as well as by Dearing (1996), applied to A level and GCSE data. For a number of years, results from applying this method to Scottish Highers were published annually as 'correction factors' by the Scottish Qualifications Authority (Sparkes, 2000, p. 178). Sparkes (2000) also used this technique in an analysis of Scottish data.

Analysis of variance

This method is described by Nuttall *et al*. (1974) as 'the most versatile and the most likely to yield sensible results with small samples and low numbers of subjects attempted' (p. 50). It essentially applies a two-way analysis of variance with unequal numbers of observations in the cells, equal weights and interaction (Scheffé, 1959). It has also been used by Backhouse (1978) and the calculation is explained in detail by Backhouse (1972, p. 140). Despite the advantages claimed for it, it is acknowledged that it is computationally complex and no reported uses of the method have been found since the 1970s.

The analysis of variance (ANOVA) method seeks to model the performance of a particular candidate in a particular examination as the sum of two factors: the ability of the candidate and the difficulty of the subject. Like Kelly's method, the ANOVA method has the advantage of taking into account simultaneously the abilities of candidates and the difficulties of the subjects they take. However, Kelly (1976a, p. 43) states that the ANOVA method is considerably more complex than other methods whose results are similar. In fact, Kelly's own method is also quite complex, and it is not obvious that there is a lot to choose between them on this score.

Average marks scaling

This method has been used in a number of Australian states for producing aggregated marks from different subjects with different 'difficulties'. Although it does not appear to have been used in the UK, and is therefore strictly speaking outside the scope of this book, average marks scaling (AMS) has qualities that make it conceptually interesting and so worth considering in this context. This method differs from the others described so far in that it aims not just to quantify the different difficulties of different subjects, but to rescale their marks onto a common scale, taking account of the spread of abilities of candidates in that subject.

Average marks scaling can be thought of as a more sophisticated version of methods such as the UBMT, ANOVA or Kelly's method. It has been applied directly to marks rather than grades, as that is how examination results are generally reported in Australia. Average marks scaling corrects both the average mark for a subject and the spread of marks, while preserving the shape of the distribution. Average marks scaling could equally well be applied to grades, provided they were coded on a numerical scale. It would then be essentially similar to Kelly's method, but has the advantage that one does not have to assume that the gaps between grades are the same in different subjects; if grades in one subject are relatively compressed, AMS will stretch them out, as well as moving them up or down. However, part of the 'interval' assumption remains as one must still assume that the gaps between grades within each subject are equal.

Average marks scaling was introduced in Western Australia in 1998, replacing a system in which similar corrections were made on the basis of scores on a reference test, the Australian Scaling Test. It was found that rescaling marks based on the average scores each student had achieved in all their subjects gave results very similar to rescaling based on the reference test, but without the need to sit an additional test (WACC, 1998).

Marks rescaled by AMS have the following properties (Partis, 1997):

- Within a subject, the order of marks and the shape of the distribution are preserved (i.e. a linear transformation is applied to each).
- The mean scaled score in each subject is equal to the mean scaled score across all subjects taken by all the students in that subject.
- The standard deviation of the scaled marks in each subject is equal to the standard deviation of the unscaled standardised marks across all subjects taken by all students in that subject.

The following heuristic account of the process of AMS draws on Partis (1997) and Seneta (1987).

1. Begin with a raw mark, G_{ni}, for each student, *n*, in each subject, *i*.
2. In each subject, standardise marks (Z_{ni}) to have mean 0 and standard deviation 1.
3. For each subject, *i*, calculate for the students who have taken that subject, the standard deviation, R_i, of all their standardized marks (Z_{nj}) in all their subjects. R_i is therefore an index of the relative spread of 'ability' of the students who have taken subject *i*.
4. For each subject, *i*, multiply the standardised mark (Z_{ni}) by R_i to get the 'range adjusted' mark U_{ni}. The standard deviations of the U_{ni} in each subject now reflect the amount of variation in ability (as measured by performance in their other subjects) of the students who have taken that subject.
5. For each subject, *i*, an adjustment, D_i, must be added to these 'range adjusted' marks (U_{ni}) to get the fully scaled marks X_{ni}. In other words, $X_{ni} = U_{ni} + D_i$. The values of the D_is are unknown at this stage.

6. For each subject, i, however, D_i will be the mean, for the students who have taken that subject, of all their scaled marks X_{nj} in all their subjects. Hence if there are n subjects, we have n equations and need to find n unknown D_is, so the D_is can be solved uniquely.

Full equations for the solution of this model are given in the appendix to this chapter.

3.3 Latent trait models

An approach that is conceptually quite different from those so far described is to base a comparison among subjects on the idea that they all measure (at least to some extent) a common trait, such as 'general ability'. In fact, references to such an idea can be found in some applications of the methods presented above. For example, in defending her approach against the criticism that different examinations may not all relate to the same kind of ability, Kelly (1976b, p. 26) conducts a factor analysis to show that they broadly do. Only one reported example of an explicitly latent trait model approach has been found, however, by Coe (forthcoming) who applied the Rasch model to GCSE data from England.

Rasch modelling

The Rasch model (Rasch, 1960/1980; Wright & Stone, 1979) provides a method for calibrating ordinal data onto an interval scale. Rasch assumes that the 'difficulty' of items and the 'ability' of persons[4] can be measured on the same scale, and that the probability of a person achieving success on a particular item is entirely determined by the difference between their ability and the difficulty of the item. In the Rasch model, these two are related by the logit function, the difference being equal to the log of the odds, and item difficulties and person abilities are estimated in logit units. Rasch's claim to provide an interval scale rests on the fact that the same difference between item difficulty and person ability anywhere on the scale corresponds to the same probability of success. For any two items of different difficulty, different persons will have different probabilities of success, but the odds ratio[5] for each person will be the same regardless of their ability, provided they fit the model. The equation for this model is given in the appendix.

Rasch analysis uses an iterative procedure to estimate item difficulties and person abilities for a given data set. It allows the fit of the model to be investigated and misfitting items and persons to be identified. It is a requirement of the model that items should be uni-dimensional (i.e. all measuring essentially the same thing) and discriminate appropriately (i.e. more able persons are more likely to be successful). Unlike other latent trait models, the Rasch model further requires that all items discriminate equally, in other words, the relationship between a person's ability relative to an item and their probability of success on it should be the same for all items. For persons, their relative probabilities of success on different items should be in line with those of others in the population.

The process of estimating grade difficulties and person abilities in the Rasch model is iterative. Given some estimate of the abilities of the candidates who have taken a particular subject (based on their overall performance in their other subjects), we can examine the relationship between the probability of a particular grade being achieved and the ability of the candidate. We can use some kind of maximum likelihood procedure to select a value for the difficulty of the grade that best explains this pattern of achievement. Having estimated grade difficulties in this way, we can then refine our estimates of candidates' abilities in an exactly analogous way, selecting a value for each person's ability that best explains their pattern of achievement of grades of known difficulty. The process is then repeated, each time using the latest estimates of difficulty and ability, until estimates converge.

Hence the estimate of the difficulty of a particular grade in a particular subject is based on all the candidates who have taken that subject with at least one other. The grade difficulty depends on the relative probabilities of that grade being achieved by candidates of different ability, as determined by their performance in all their subjects and taking into account the different difficulties of all the grades they have gained.

In this way the Rasch approach is quite similar to the subject 'matrix' methods described above, though it differs in two important respects. The first is that with Rasch it is possible to estimate the difficulties of each grade in each subject independently, using a 'partial credit' model (Masters, 1982). Hence there is no need to make any kind of interval assumption about the scales on which grades are coded; the Rasch model automatically assigns a value to each grade on a scale which may be said to have the 'interval' property, that is, the same interval anywhere on the scale denotes the same difference in the probabilities of being achieved. This is a potentially important advantage since to use methods such as Kelly's or ASPA we must assume not only that the intervals between different grades in the same subject are equal, but also that these intervals are the same across all subjects.[6] Given Coe's (forthcoming) finding that the intervals between grades are far from equal, this may be a significant advantage for the Rasch approach.

The other key difference is that the Rasch model requires the subjects and candidates analysed to fit a particular model. In this context, fitting the model means that it must be possible to assign ability levels to all persons and difficulty levels to all items (i.e. subjects and grades) such that when we consider all candidates of a particular level of ability who have taken a particular subject, the proportion of them who achieved a particular grade should be reasonably close to what is predicted by the model. A key requirement for such fit is that both difficulty of items and ability of persons are uni-dimensional. In other words, there must be essentially just one kind of 'ability' which persons differ in the amount of which they exhibit and which largely accounts for their performance at all grades in all subjects.

If a particular subject, or at least a particular grade in a particular subject, does not fit the model, this will be evident and we can see which grades are not behaving as 'expected'. We can also identify any individuals or groups of candidates who are 'misfits'. The fact that the Rasch model specifically requires a uni-dimensional concept of ability seems to lend its results to interpretation in terms of the general level of ability represented by a particular achievement.

4 Criticisms of common examinee approaches

A number of writers have discussed issues arising from the use of these methods, including Christie & Forrest (1981), Newbould (1982), Forrest & Vickerman (1982), Alton & Pearson (1996), Pollitt (1996), Cresswell (1996), Goldstein & Cresswell (1996), Fitz-Gibbon & Vincent (1997), Newton (1997; 2005) and Jones (2003).

The main criticisms of statistical comparisons may be listed under five headings, and are summarised briefly below. For now, these criticisms are simply stated as they have been presented by their authors. I return to evaluate these criticisms in a later section, having first tried to clarify some of the assumptions required by different interpretations of the methods and the different conceptions of 'comparability' on which they rest.

4.1 Factors other than 'difficulty'

This criticism has already been mentioned above in suggesting that better teaching in English, or higher levels of motivation, might account for candidates' better performance in that subject.

A number of writers point out that examination performance is affected by many factors apart from 'difficulty', so, unless we are prepared to assume (or can show) that these factors are equal or unimportant, we cannot judge 'difficulty' simply by comparing outcomes. Just because common candidates typically get lower grades in one subject than in others it does not necessarily follow that it is more difficult. A number of other factors, including the intrinsic interest of the subject, the quality of teaching experienced, extrinsic motivations such as the need for a particular qualification, the candidates' levels of exam preparation, the amount of curriculum time devoted to it, and so on, could all affect performance, without making that subject more 'difficult' (Alton & Pearson, 1996; Goldstein & Cresswell, 1996; Newton, 1997).

4.2 Multidimensionality

In order to compare standards in different subjects we have to have some basis for comparing them. This amounts to saying that different subjects must all measure the same thing, or at least have some significant trait in common. In other words, that subjects being compared must be uni-dimensional – which of course they are not. It is meaningless to say, for example, that 'art is easier than physics'; they are just different. Goldstein & Cresswell (1996) give an example of a comparison between a spelling test and a degree in English; in theory one could make the two equal in difficulty, at least in a statistical sense, though it would be absurd to say that they were equivalent.

Another subtle variation on this argument is provided by Pollitt (1996). He gives an example of a set of examination results in different subjects where five candidates of equal 'general ability' choose different subjects in which they have different specific aptitudes. When these different aptitudes in different subjects are allowed to interact with different reasons for choosing them, Pollitt shows that the illusion of differential difficulty is created. Some subjects (English and economics in his example) are chosen by those who are best in those subjects, while others (mathematics) are chosen by those who are relatively weak in that subject. The result is that candidates typically do worse in mathematics than in their other subjects and hence mathematics appears 'harder'.

4.3 Unrepresentativeness

The groups of students taking particular combinations of subjects (on whom statistical comparisons are based) are not representative of all those who take (or might take) a particular subject. Again, an example of this has already been mentioned: students who take both physics and media studies are unlikely to be representative of those taking either subject. This point is made by, for example, Goldstein & Cresswell (1996).

Newton (1997) discusses the question of exactly who it is any group of candidates on whom a statistical comparison is made should be representative of. In other words, to what population do we want any claims of comparability to apply? He argues that this should be the whole population of students in the cohort, whether or not they choose actually to take a particular subject. In this case, unless we have truly representative (e.g. random) samples taking every subject, any claims about comparability are very problematic.

4.4 Sub-group differences

If we analyse subject 'difficulties' for different sub-groups (e.g. males and females) we get quite different results. For example, for males, history may appear 'harder' than mathematics, while for females mathematics is the 'harder' (Pollitt, 1996). We might also find that for candidates who take mathematics with it, physics is really no harder than any other subject, whereas for those who take it without mathematics, physics appears substantially more difficult (Rutter, 1994). Hence a judgement about whether one subject is 'harder' than another depends very much on who happened to take those subjects. And if the characteristics of the entry change, so would the supposed 'difficulties' (Alton & Pearson, 1996; Pollitt, 1996; Newton, 1997; Sparkes, 2000). The existence of different relative difficulties for different sub-groups is also a challenge to the assumption of uni-dimensionality.

4.5 Problems of forcing equality

Adjusting the 'difficulties' of different subjects to make them all equivalent would cause problems during any changeover period for users, the public and professional bodies. Some of the currently 'harder' subjects would need to have absurdly high pass rates, while subjects currently graded 'leniently' would have to be failed by most candidates. This situation would be satisfactory for neither group. Requiring grade boundaries to be modified in this way would change the nature of the examining process and could delay the publication of results. There is also the problem that different methods of estimating relative 'difficulties' would give different corrections, and there is no clear consensus about which method is best (Alton & Pearson, 1996; Goldstein & Cresswell, 1996).

5 Clarification of assumptions underlying the methods, and different conceptions of 'comparability'

The criticisms presented above represent a formidable challenge to the validity of common examinee methods. Before we can evaluate these criticisms, however, we must be clear about exactly what assumptions are required by the different methods, and, in particular, by different interpretations of their results. This leads us to an argument that different uses of these methods may rest upon quite different conceptions of what is actually meant by 'comparability', and that some of the criticisms of these methods are based on different conceptions again.

5.1 Assumptions required by the methods

Attempts to clarify the assumptions that underlie the use of common examinee methods seem to have come more often from their critics than their proponents. Some of those who have used these methods have perhaps been less critical of them and less concerned with justifying and validating than with reporting the results (e.g. Fitz-Gibbon & Vincent, 1994). The critics, on the other hand, have often discussed in some detail the validity of the assumptions claimed to underlie the method, for example, the requirement that factors such as motivation in a particular subject, or the quality of teaching experienced, can be treated as equal, or that candidates taking a particular pair of subjects are representative of all those who might take either of them, or that a group of subjects being compared is sufficiently 'uni-dimensional' to justify a basis for comparability (e.g. Forrest & Vickerman, 1982; Newton, 1997).

However, the question of what assumptions are required by a model cannot be answered without considering the interpretations and uses to which any results are put. In fact one could argue that any statistical model, including common examinee methods, can be applied to any data set without strictly speaking having to make any assumptions at all. It is only when one comes to trying to interpret the results that it is really appropriate to talk about validity. An interpretation may be valid or invalid, but a method, without a specific interpretation attached, cannot. In general, the application of any statistical method can be interpreted in a number of ways. One particular interpretation of the results might be valid only if a certain set of assumptions are made, whereas another interpretation could require an entirely different set.

Hence, before we can consider what assumptions are required by these models, we must first be clear what kinds of interpretations might be made of the results. Before we can do this, we must first be clear about the meaning of 'comparability' in this context, or, rather, about the different, but not always well separated, meanings that word can have.

5.2 Conceptions of comparability

I have argued elsewhere in this volume[7] that there are three broad, distinct conceptions of comparability as applied to examination standards. It is important to separate these different meanings of the word 'comparability' as once the distinction is made it becomes clear that much of the discussion of the problems of common examinee methods is actually at cross-purposes. Different arguments (or sometimes different parts of the same argument) have made different interpretations of concepts like 'difficulty' or 'standards', based on different fundamental conceptions of 'comparability'. While it is broadly true that a particular view of the meaning of 'comparability' implies certain interpretations of such concepts, there is not such a simple relationship between these different conceptions and specific methods, such as SPA.

The three conceptions of comparability are *performance* comparability, *statistical* comparability and *construct* comparability. Each of these is briefly outlined below, together with their associated interpretations of terms such as 'difficulty' and 'standards', and examples of applications of common examinee methods that have drawn on each. The extent to which each conception addresses the problem of different relative difficulties for different sub-groups is also discussed, as is the question of which subjects can be compared under each view.

Performance comparability

According to a *performance* view of comparability, the 'standard' of a particular award resides in the levels of skill, knowledge, understanding – or any other qualities – that are required to achieve it. One examination would be seen as more 'difficult' than another if it required skills, knowledge or understanding that were more advanced, in other words if it made a greater demand on the candidate.

This conception often appears to be the default in thinking about comparability, though it is not often explicitly stated. When writers do not explicitly attempt to define concepts such as 'difficulty' or 'standards' it often seems to be implicit that they are thinking in terms of the intellectual demands made by an examination, the skills, knowledge and understanding that must be demonstrated to gain the award of a particular level. However, it is hard to find a clear illustration of this perspective and no doubt many writers who appear to adopt it would object to being so classified.

According to a purely *performance* view of comparability, it should be possible to specify the difficulty of an examination even if no candidate has ever taken it. Perhaps for this reason, the existence of sub-group differences, as illustrated by the example of mathematics and English GCSE, above, has been taken by some critics of common examinee methods as an unequivocal death blow to any hopes of establishing comparability by statistical means (e.g. Newton, 1997). The implicit assumption here may be that all comparability is essentially *performance*. However, other conceptions of comparability may have less difficulty with the problem of sub-group differences.

Another issue that is influenced by our perspective on comparability is which subjects can legitimately be compared. From a *performance* view, only examinations that give rise to the same phenomena can be compared; there must be common skills, knowledge or understandings, defined in terms of common criteria. Discussion of comparability is often presented in terms of identifying 'cognate' subjects; in other words, subjects that are similar in terms of their disciplinary roots, content area or methods. This idea seems to imply a *performance* perspective, since one could only judge the relative demand of tasks that have these kinds of similarities.

Statistical comparability

The second type, *statistical* comparability, holds that two examinations may be seen as comparable if a 'typical' candidate has an equal chance of achieving a particular level in each. Under a *statistical* conception of comparability, the 'standard' depends on its likelihood of being reached, possibly after taking into account other factors. An examination level is 'harder' if it is rarer, or at least estimated to be less likely to be achieved by a 'similar' candidate. Different operationalisations of this general approach include simple norm (cohort) referencing, the use of value-added models (multilevel or otherwise) and, of course, common examinee methods. However, we must be clear that the method itself does not necessarily imply a particular view of comparability; it depends how the results are interpreted.

Examples of *statistical* conceptions of comparability in the use of common examinee methods are perhaps a little easier to find than *performance* ones, though again they are seldom unequivocal. The statement by Nuttall *et al*. (1974) seems to adopt this perspective.

> … we can see no logical reason why, if a large group of candidates representative of the population took, for example, both English and mathematics, their average grades should not be the same.
>
> Nuttall *et al.* (1974, p. 12)

A similar view can also be found in Fitz-Gibbon & Vincent (1994).

> The term 'difficult' cannot be taken as meaning necessarily or intrinsically difficult. Rather, subjects are said to be either 'difficult' or 'severely graded' if the grades awarded are generally lower than might have been reasonably expected on the basis of adequate statistics.
>
> Fitz-Gibbon & Vincent (1994, p. i)

This definition suggests that there may actually have been no real disagreement between Fitz-Gibbon & Vincent (1994; 1997) and Goldstein & Cresswell (1996); they were simply using the same word to mean two quite different things.

One feature of the *statistical* conception is that it seems to have less difficulty with the problem of sub-group differences. On average, candidates who take both subjects may have a better chance of success in English than they do in mathematics. However, if we know that a candidate is male, for example, our estimate of their relative chances may change. From a *statistical* viewpoint there is no particular reason why an estimate of chances of success should not depend on the characteristics of the candidate as well as on the subjects taken.

While it might be considered a desirable characteristic of different examinations that their relative difficulties should be the same for, say, males and females, it is not a pre-requisite for an understanding of *statistical* comparability between them. Of course, if it turns out that relative difficulties are indeed quite different for different sub-groups, then we can no longer talk about comparability of subjects per se but only of subjects in relation to particular sub-groups.

The *statistical* conception of comparability is the most broad-minded on the question of which subjects can be compared. If comparability is based on the concept of chances of success then there seems to be no reason why any groups of subjects cannot be compared. This conception makes no requirement for different subjects to be related in any way, only that a particular level of achievement should be equally rare in each.

Construct comparability

The third type, *construct* comparability, holds that two examinations may be compared if they have some construct in common. For this version of comparability, the 'standard' of a particular examination performance depends on the level of the linking construct that it signifies. One examination is 'harder' than another if it indicates a higher level of the linking construct.

An example of this kind of comparability can be found in Fitz-Gibbon & Vincent (1997) who talk about the 'common currency' of A level grades. By this they mean that for some purposes, such as when admissions tutors in UK universities make decisions about which applicants to accept, grades in different subjects may be treated as interchangeable.

> What our analyses suggest is that the 'common currency', i.e. that which can be seen as the information contained in any grade about general aptitudes, can be better operationalised by recognising differences between the subjects in 'difficulty'.
>
> Fitz-Gibbon & Vincent (1997, pp. 293–4)

Here the linking construct is 'general aptitudes', though other linking constructs could be imagined. In the context of an admissions tutor using grades in a subject other than their own to infer a candidate's suitability for entry, we might speculate that the construct of interest would be that student's generalisable capacity for learning in another academic context, a construct that is probably reasonably well summarised by the term 'general aptitudes'.

If it is accepted that all the subjects being compared measure (at least to some extent) 'general aptitudes', then we can legitimately compare their outcomes. If we do compare them, then we must interpret these comparisons in terms of our construct of 'general aptitudes'. So in saying that, for example, physics is 'harder' than biology we mean that a particular grade in physics indicates a higher level of 'general aptitudes' than would the same grade in biology.

Another context mentioned by Fitz-Gibbon & Vincent (1997) is the use of examination grades in school performance tables. Here the grades, perhaps after adjustment for the effects of prior attainment or other factors, might be taken as an indication of the effectiveness of the teaching received, so we have an alternative linking construct from the same analysis and with it the possibility of an entirely different interpretation of the differences in 'difficulty' that were found. Indeed, the same study was cited above as exemplifying a *statistical* conception of comparability, so it is clear that the same method can support more than one interpretation.

The problem of sub-group differences is not entirely solved by adopting the *construct* conception of comparability. Some level of uni-dimensionality is required by the assumption of a linking construct and significant variation in relative 'difficulties' for different sub-groups would undermine this. There are a number of possible ways to get around this problem, though none of them is really completely satisfactory.

One approach would be to limit any comparison to groups of subjects in which there were no substantial sub-group differences. For example, we might say that mathematics and English cannot really be compared in relation to a common construct because they are not sufficiently uni-dimensional. We cannot infer levels of 'general aptitudes' from performance in these two subjects since the aptitudes required by each appear to be too specific. On the other hand, despite the differences in the comparison of mathematics and English for different sexes, grades in these two subjects are highly correlated (r = 0.77 in the 2004 data set) so in this particular case we might well conclude that they are sufficiently uni-dimensional for our purposes.

Another approach would be to limit comparisons to particular sub-groups, so we could compare mathematics and English separately for males and females, and accept that 'comparability' will be gender-specific. However, this requires us to invoke gender-specific linking constructs such as 'male general aptitudes' and 'female general aptitudes' – a rather strange idea.

The question of which subjects can be compared is of course related to this issue. If our interpretation of differences between subjects draws on the notion of a linking construct such as 'general aptitudes' to provide a basis for comparability, then comparisons must be limited to subject examinations that broadly measure that trait. Achievements in any subjects being compared would therefore have to correlate reasonably well with each other, so we might adopt a largely empirical criterion to decide whether subjects are comparable or not.

Note that this last criterion is likely to contrast with the idea of 'cognate' subjects since it would be quite possible for examinations even with the same subject title to correlate very poorly. An example of this might be found at GCSE where under the heading of 'science' we would find biology, chemistry and physics, along with combined science, but also vocational science. Variations in the modes of assessment used could make more difference to the correlations among syllabuses than their nominal content.

6 Defence of common examinee methods

Given the number of accounts of uses of the common examinee methods that have been outlined above, and the force of the criticisms that have been made against them, there is surprisingly little in the existing literature by way of systematic defence of these methods. One exception is Fitz-Gibbon & Vincent (1997), who provide a specific reply to the critique of Goldstein & Cresswell (1996). Another is Kelly (1976b) who addresses specific criticisms made of an earlier conference paper she gave. As with the criticisms, however, much of the defence of these methods has been unclear about the exact meanings of terms such as 'difficulty', or, if it has been clear, has not been consistent.

6.1 Evaluation of these criticisms in relation to specific interpretations

I now return to the five broad criticisms of common examinee methods and attempt to evaluate them.

Factors other than difficulty

Viewed from a *performance* perspective, it must be acknowledged that this criticism seems to have some weight. If one interprets the 'difficulty' of an examination as a function of its level of demand in terms of specific criteria, then it is hard to deny that a statistical difference between subjects does not necessarily indicate a difference in 'difficulty'. Given all the other factors that may affect attainment, it seems unsatisfactory either to ignore them, or to assume they will be the same for all subjects. In theory, we could try to measure and take account of them all, but in practice this would also be likely to be somewhat unsatisfactory.

On the other hand, if our conception of comparability is fundamentally *statistical*, then we are likely to interpret the kinds of differences in attainment revealed by common examinee methods as indicating different chances of success in different subjects. From this perspective, it may not matter that the reasons for these differences may vary, since whatever the reasons, the differences remain. This is particularly the case if it is judged to be perfectly acceptable that a typical candidate's chances of success differ in different subjects. However, even if differences are held to be unacceptable on the grounds that one's chances of success should be the same regardless of subject, then one could still attempt to compensate for the 'other factors' so that an individual candidate could make a fair choice between them based on equal expectations.

Similarly, from a *construct* comparability perspective, differences in attainment by equivalent candidates are taken to indicate different conversion rates between a grade in a particular subject and some underlying linking construct such as 'general aptitude'. The example given above, of the admissions tutor judging the suitability of candidates with results in different subjects, illustrates that there may be situations in which any other factor that affects attainment could be seen as wholly irrelevant.

In considering this criticism generally, it is interesting to note that most of the writers who cite the problem of 'other factors' list a number of factors that might in principle affect performance, without actually providing any evidence that these factors do in fact vary by subject. An exception is Newbould (1982) who shows that differences in the apparent severity of grading of different subjects agree reasonably closely with students' expressed preferences for those subjects. In other words, the subjects that appear hardest are also generally the least liked. Newbould speculates that differences in performance levels in different subjects might be accounted for by these differences in preferences. However, liking for a subject is not the same as motivation to work in it and a correlation, no matter how strong, does not necessarily imply that relative attainment is a result of differential preference. It could equally be, for example, that the causal relationship could be the other way: students like most those subjects in which they are doing best.

Multidimensionality

Most critics of statistical approaches to comparability point to the assumption of uni-dimensionality as an untenable requirement. Clearly, examinations in different subjects measure different things; otherwise there would be no point in having different examinations. Hence the basis on which a comparison can be made is problematic. This issue has been discussed briefly above in relation to the restrictions on which subjects can be compared according to the three different conceptions of comparability.

From a *performance* perspective, multidimensionality is a real problem. According to this conception of comparability, we can really only compare two examinations against the same criteria. They may be comparable only to the extent that common phenomena may be observed as the output of both; if there is no such common ground, the question of which is 'harder' is meaningless.

From a *statistical* conception this issue is much less of a problem; indeed we may even deny that any kind of uni-dimensionality is actually a requirement. For example, if we define 'standards' in terms of population norms, there is no requirement that different standards should be measuring the same thing, only that they should be equally rare. Although this kind of cohort referencing is not a feature of common examinee methods, the same logic might be applied to them. For example, the results from Kelly's method could be interpreted as indicating the relative chances of success in different subjects of candidates who are typical of those who actually entered them. An analysis might show, say, that a vocational course in information technology was 'easier' than a traditional examination in Latin. In this context, 'easier' just means that candidates who were similar in terms of their achievement in their other subjects typically achieved better grades in the former; there is no assumption that these two examinations are measuring the same thing.

The position of the *construct* conception of comparability is perhaps somewhere between the other two. If we are to invoke the concept of a linking construct, we must maintain that different subjects, while not perfectly uni-dimensional, are at least sufficiently uni-dimensional to allow meaningful comparison. A pragmatic restriction here would be to accept that making comparisons across all subjects is going too far but that there are groups of 'cognate' subjects, within which a comparison may be valid.

In this context, 'uni-dimensional' may be taken to mean that a large part of the variation in grades in different subjects is shared. Measures such as the correlations between grades in different subjects, loadings on a single factor in a factor analysis, or measures of internal consistency such as Cronbach's Alpha could all provide evidence of this. The Rasch model has a particular requirement for uni-dimensionality, since if this assumption is violated to a large extent the invariance properties of the model (i.e. item-free measurement) are lost. When this happens the results may be misleading or not usable, though in practice some tolerance for deviation from strict uni-dimensionality may be not just necessary, but desirable for valid interpretation of the underlying latent construct (Hambleton, 1989; Linacre, 1998).

The issue of uni-dimensionality is therefore essentially an empirical one; for any set of subjects we can calculate the extent to which they overlap. However, there is no absolute threshold at which a set of subjects is clearly uni-dimensional. Any such threshold would be arbitrary and the question is really one of degree rather than kind. The decision about whether they are uni-dimensional enough must be a matter of judgement.

Having said that, given any reasonable minimum threshold for the amount of overlap that a set of subjects must exhibit in order to be accepted as uni-dimensional, it is beyond question that such a set exists. The higher the correlation we require among different subjects, the smaller the group is likely to be. It is also likely, however, that some subjects could not be included in the group without having to set an unacceptably low threshold. Hence the question is not whether all subjects can be considered uni-dimensional, but which subjects can reasonably be considered sufficiently uni-dimensional. An example of this can be seen in Coe's (forthcoming) application of the Rasch model. He judged that 34 GCSE subjects were sufficiently uni-dimensional to allow them to be compared, but that many others (including General National Vocational Qualifications and some creative GCSE subjects like music and art) had to be excluded in order to meet this requirement.

Unrepresentativeness

This criticism is usually targeted at subject pairs analysis (SPA). For example, Goldstein & Cresswell (1996, p. 438) point out that 'students who happen to take particular pairs (or combinations) of subjects are not typical of either subject'. In fact, many uses of subject pairs (e.g. Nuttall, *et al*. 1974) estimate the severity of a subject's grading by calculating an average of the subject pair differences across all subjects taken with it, so are really ASPA (see above). Hence, the estimate of the severity of, say, chemistry is based on all the students who have taken chemistry and at least one other subject, which in practice is pretty close to being all those who have taken chemistry at all. Moreover, most other common examinee methods, such as that used by Kelly (1976a), or the Rasch model, also base their estimates of a subject's difficulty on all students who have taken it with at least one other. Hence the objection as it is often presented is unfounded.

However, there is a bigger problem here. In certain subjects, the students who typically take them are severely unrepresentative of all those who might potentially take them, and this makes comparisons extremely problematic. In the context of England, for example, examinations in languages such as Urdu or Chinese are taken disproportionately by native speakers of those languages. For a native speaker, a GCSE in Urdu is likely to be easier than the other GCSEs they take and hence they are likely to achieve higher grades on average. This is likely to result in statistical comparisons showing Urdu to be an easy subject (see, for example, Yellis, 2006). This would be no reflection of the difficulty that a non-native speaker would experience to achieve the same.

From a *performance* perspective, this would be a significant problem. However, even the *statistical* conception has something of a problem here, since a person's chances of success would depend significantly on whether they were a native speaker. Hence we would have to add a rider to our interpretation and say that the differences in achievement indicate different chances of success *for students typical of those who typically take the subject*. If a particular student is different in some way from those who typically take that subject, or indeed if they are not known to be the same, then an estimate of their chances of success should be treated with some caution. However, this is not to say that is wholly worthless, since a sensible person will treat most things with some caution anyway.

From a *construct* view the difficulty seems a little less. Although we would admittedly make an incorrect judgement about the 'general aptitude' of an atypical entrant based on their result in that subject, the evidence from that subject is likely to be a small part of the total evidence available. If a judgement is being made about an individual student then it is also possible that their atypicality could be made known.

Sub-group differences

A number of analyses of subject difficulties (e.g. Massey, 1981; Newton, 1997; Sparkes, 2000) have reported that if apparent differences between subjects are found, they vary considerably when the analysis is limited to particular sub-groups. As has already been stated, such variations may be problematic for the *performance* conception of comparability, since 'difficulty' ought to be independent of the particular population of candidates. The issue of sub-group differences presents a particular problem if it is argued that differences in difficulty are inherently undesirable, or that grades should be corrected to eliminate them.

However, from a *statistical* conception, such inconsistencies may be less problematic. Our estimate of a person's chances of success is likely to depend on our knowledge of their characteristics. The factors that have been shown to affect the 'difficulties' of subjects, such as gender, type of school, prior attainment or subject combinations (Newton, 1997; Sparkes, 2000) might all be expected to affect a person's relative chances of success in different subjects. There is no contradiction in saying, for example, that overall, candidates have a better chance of doing well in English than they do in mathematics, but that for boys the reverse is generally true. We can estimate a person's chances of success even if we know nothing about that person, but knowledge of their characteristics will enable us to make a better estimate.

Similarly, within a *construct* perspective, our estimate of a person's 'general aptitude' may also be modified by knowledge of the particular sub-groups to which they belong. It would be possible, for example, initially to estimate a person's aptitude as high, but then to find that they had entered a combination of subjects in which candidates with their characteristics tended to do relatively well, and hence to have to revise our estimate downwards. Hence, provided we remember that statistical differences in performance tell us something about the candidates as well as about the subjects, the problem of sub-group differences may not really be such a problem at all.

One way to interpret these sub-group differences is in terms of bias in the examination process. For example, if history appears harder for males than it does for females, compared to other subjects (as in Pollitt's, 1996, example), this could be seen as indicating that the examination process in history is biased towards females (relative to mathematics). Of course, the 'examination process' here must include, as well as the examination itself and its assessment procedures, the teaching and learning that preceded it. The 'bias' could arise from factors such as differential motivation or effort, better teaching or examination preparation. It could also arise from selection processes or choices about who enters different subjects. In these cases the word 'bias' might not really be appropriate, so it may be important to try to understand the reasons for any differences in relative performance.

Problems of forcing equality

Even if one interpreted performance differences as indicating subject difficulty, it would be perfectly possible to believe that some subjects were harder than others, but that there was no reason to try to change this. Indeed, this appears to be the official stance taken in Western Australia (WACC, 1998). Hence the position that all subjects should be forced to be equivalent is not necessarily implied even by a simplistic interpretation of statistical differences.

However, if one takes the view that standards across subjects should be aligned, then the question remains how this should be done. The rejection of statistical approaches offers no real solution here, since one can no more *judge* the intellectual demand of an assessed attainment than calculate it statistically. Of course it would be possible to align the standards for the population as a whole, but consideration of sub-groups might show substantial misalignment. Sub-groups themselves could then also be aligned, artificially equating the scores of, for example, males and females, though in practice it would be hard to guarantee that all possible sub-groups had been considered. Ensuring that standards are comparable across subjects is far from straightforward (Baird *et al.*, 2000).

Accepting that different subjects may offer different chances of success does also have its problems, but these can be resolved. The main difficulties arise when examination grades are taken as an indication of something other than simply achievement in a particular course of study. For example, if grades are used as a selection tool to indicate a person's general academic ability, or aggregated into league tables to denote the quality of teaching provided in a school, then it matters if some subjects are more severely graded than others. However, if grades are to be used in this way, then it ought to be possible to calculate a fair exchange rate to equate and convert them into some kind of currency fit for these purposes. This in fact is exactly what is done in Western Australia.

Whether we wish to equate standards or merely to calculate an exchange rate between them, however, we are still left with the problem of which method to choose. A number of studies have claimed that this choice makes a substantial difference to the outcome (e.g. Alton and Pearson, 1996). Others, on the other hand, have claimed that there is broad agreement among the different methods (e.g. Nuttall *et al.*, 1974). This debate seems hard to resolve on the basis of existing evidence. And it may be important to distinguish between the amount of disagreement among different methods for estimating subject 'difficulties' and the scale of the differences among subjects, by whatever method.

7 Conclusions: strengths and weaknesses of common examinee methods

7.1 The different methods compared

Whether conclusions from any of these methods are judged to be valid and, if so, conclusions from which of them are judged to be best must depend on the purpose of the comparison. Unless we are clear about why we want to compare different subjects, we cannot really advocate any particular method for doing so. For this reason, it has been important to try to clarify the different bases on which comparisons might be made and how they may be interpreted.

If we adopt a *performance* conception of comparability, then it seems that none of the statistical methods can really do this. The 'difficulty' of a task is so dependent on contextual factors beyond the control of the examination process that it becomes very hard to estimate, either by statistical or judgement methods.

Alternatively, if our perspective is *statistical* then our intention may be simply to compare how likely the same grade is to be achieved. In this case, several of the approaches may be appropriate. Many of the variations of the basic SPA methods, along with Kelly's method or the AMS approaches appear to offer satisfactory solutions to this problem. Of these, AMS is perhaps the best, given its flexibility, sophistication and ease of calculation.

Of course, we must remember that different students with different characteristics will have different probabilities of achieving the same grades, so these likelihoods are not absolute. They are meaningful only when applied to a particular group. Given that a number of different methods appear to be appropriate, it seems important to know how similar their results would be. This does not seem to be clear from our existing knowledge.

This interpretation of differences in grades achieved implied by the *statistical* conception of comparability (that they indicate different chances of success) may be seen to underpin their use in accountability processes such as league tables. Although examinations such as GCSE and A level may never have been intended to be used in this way, such uses are a significant part of their function today. If we take an integrated view of validity (Messick, 1989), then we cannot ignore this issue of consequential validity in considering whether grades are awarded appropriately. If the same grade is more likely to be achieved by a typical candidate in one subject than another, yet both grades are awarded the same value in an accountability process, then it does seem likely that problems will result.

A third rationale for comparison has also been considered: *construct* comparability. Here, in principle, the same grade in different subjects ought to indicate the same level of a common trait, general academic ability. For this interpretation the Rasch model may be the most appropriate, since it explicitly compares different subjects on the basis of a single dimension. The use of ASPA weighted by inter-subject correlations might also address this interpretation. AMS approaches are used for this purpose in Australia, so presumably this and the other approaches that are conceptually similar (subject pairs and Kelly's method) could also be used.

This interpretation of grades in terms of 'general aptitude' is important in practice since it is implied by the use of examination grades for selection into further or higher education – a widespread practice with significant consequences attached to it. If grades are to be used in this way then it is important that the same grade in different subjects should denote the same level of ability. One limitation of the Rasch approach here is that not all subjects can be compared; only those that are sufficiently aligned with the unitary ability construct can be included. However, if examinations depend on substantially different skills it probably is right that they should not be directly compared. Achievements in any examinations that do not fit the uni-dimensional model tell us nothing about that candidate's 'general aptitude' and should not be treated as comparable to those that do fit.

7.2 Directions for further research

A number of areas for further research have been identified already. For example, there is debate about how well the results of different methods agree. More research on this question would be useful to resolve this. Another issue that has not been fully investigated is whether other factors (such as motivation, quality of teaching or preparation, etc.) can really account for the differences we see in attainment. If so, it might be possible to retain the *performance* definition of comparability, but use statistical methods to investigate it.

8 Conclusion

It is clear that no statistical process, however sophisticated, can provide a full and satisfactory answer to the problem of evaluating the comparability of different examinations. There are too many different meanings of the word 'comparability', too many unknown but potentially important confounding factors, and too many different ways in which examinations are used. We must therefore seek to be clearer about our underlying conceptions of comparability and about the specific purposes to which we wish to put examination results. We must also be cautious about the results from any single method, and seek to triangulate the evidence from different approaches involving different assumptions and methods. We must be careful, though, not to combine or make direct comparisons between results that arise from incommensurable conceptions of 'comparability'.

However, from the point of view of monitoring 'standards' in the UK, it does seem clear that there are differences in the grades that typical candidates might expect to achieve in different subjects. It is also clear that the same grade in different subjects can indicate quite different levels of underlying 'aptitude'. Given the ways examination grades are used for purposes such as accountability and selection, these differences give rise to anomalies, the consequences of which can be quite serious. There may be a case, therefore, for the methods that have been outlined in this chapter to play, and be seen to play, a bigger part than they currently do within England in the processes of monitoring and ensuring 'comparability'.

Endnotes

[1] This example and the discussion of its interpretation draws on a similar example presented by Newton (1997).

[2] Note that most of the early uses of this method were limited to within-board datasets. Later national analyses, such as by Willmott (1995), were possible only because datasets matched across boards started to become available.

[3] This result is based on the average differences in the grades achieved during the period 1994–2003, adjusting for the differences in ability of the candidates in different subjects using Kelly's method.

[4] The words 'difficulty' and 'ability' are used generally in discussing the Rasch model, even when their normal meanings are considerably stretched. For example, in the context of a Likert scale attitude item one may talk about the 'difficulty' of an item to mean its tendency to be disagreed with (i.e. how 'hard' it is to agree with). The use of these words may initially seem strange to anyone not

familiar with the Rasch model. However, I have adopted this convention, partly in order to comply with standard practice, and partly because although the words 'difficulty' and 'ability' are not quite right for the interpretation intended, I am unable to think of better ones.

5 The odds ratio is the ratio of the odds of the two probabilities. In other words if a person has probabilities p and q of success on two items, the odds are $p/(1 – p)$ and $q/(1 – q)$ respectively. Hence the odds ratio is $[p/(1 – p)]/[q/(1 – q)]$. The logit function is:

$$\text{logit}(p) = \ln[\, p \,/\, (1 - p)]$$

so the log of the odds ratio is the same as the difference in the two logits, logit(p) – logit(q).

6 Of course we do not strictly have to assume that they are equal, but we have to make some assumption about their relative sizes. Note also that the AMS method requires an assumption about grade intervals within subjects, but not between subjects.

7 Commentary on Chapter 4.

References

Advanced Level Information System. (2004). *A level subject difficulties*. The Advanced Level Information System, Curriculum, Evaluation and Management Centre, University of Durham.

Alton, A., & Pearson, S. (1996). *Statistical approaches to inter-subject comparability*. Report for the Joint Forum for the GCSE and GCE.

Backhouse, J.K. (1972). Reliability of GCE examinations: A theoretical and empirical approach. In D.L. Nuttall & A.S. Willmott (Eds.), *British examinations: Techniques of analysis*. Slough: National Foundation for Educational Research.

Backhouse, J.K. (1978). *Comparability of grading standards in science subjects at GCE A level.* Schools Council Examinations Bulletin 39. London: Evans/Methuen.

Baird, J., Cresswell, M.J., & Newton, P. (2000). Would the *real* gold standard please step forward? *Research Papers in Education, 15*, 213–229.

Bardell, G.S., Forrest, G.M., & Shoesmith, D.J. (1978). *Comparability in GCE: A review of the boards' studies, 1964–1977*. Manchester: Joint Matriculation Board on behalf of the GCE Examining Boards.

Christie, T., & Forrest, G.M. (1981). *Defining public examination standards.* Schools Council Research Studies. London: Macmillan Education.

Coe, R. (forthcoming). Comparability of GCSE examinations in different subjects: An application of the Rasch model. *Oxford Review of Education, 34.*

Cresswell, M.J. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues* (pp. 57-84). Chichester: John Wiley and Sons.

Dearing, R. (1996). *Review of qualifications for 16–19 year olds.* London: School Curriculum and Assessment Authority.

Fearnley, A.J. (1998). *Update on an investigation of methods of analysis of subject pairs by grade*. Unpublished research paper, Northern Examinations and Assessment Board.

Fitz-Gibbon, C.T., & Vincent, L. (1994). *Candidates' performance in public examinations in mathematics and science*. A report commissioned by the School Curriculum and Assessment Authority from the Curriculum, Evaluation and Management Centre, University of Newcastle-upon-Tyne. London: School Curriculum and Assessment Authority.

Fitz-Gibbon, C.T., & Vincent, L. (1997). Difficulties regarding subject difficulties: Developing reasonable explanations for observable data. *Oxford Review of Education*, *23*, 291–298.

Forrest, G.M., & Smith, G.A. (1972). *Standards in subjects at the Ordinary level of the GCE, June 1971*. Occasional Publication 34. Manchester: Joint Matriculation Board.

Forrest, G.M., & Vickerman, C. (1982). *Standards in GCE: Subject pairs comparisons, 1972–1980*. Occasional Publication 39. Manchester: Joint Matriculation Board.

Fowles, D.E. (1998). *The translation of GCE and GCSE grades into numerical values*. Unpublished research paper, Northern Examinations and Assessment Board.

Goldstein, H., & Cresswell, M.J. (1996). The comparability of different subjects in public examinations: A theoretical and practical critique. *Oxford Review of Education*, *22*, 435–441.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: American Council on Education and Macmillan Publishing Company.

Ireson, G. (1996). The effect of studying A level mathematics on the A level physics grade achieved. *School Science Review*, *77*(280), 116–119.

Jones, B.E. (2003). *Subject pairs over time: A review of the evidence and the issues*. Unpublished research paper RC/220, Assessment and Qualifications Alliance.

Kelly, A. (1976a). A study of the comparability of external examinations in different subjects. *Research in Education*, *16*, 37–63.

Kelly, A. (1976b). *The comparability of examining standards in Scottish Certificate of Education Ordinary and Higher grade examinations*. Dalkeith: Scottish Certificate of Education Examination Board.

Linacre, J.M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, *2*(3), 266–283.

Massey, A.J. (1981). *Comparing standards between AL English and other subjects*. Test Development and Research Unit, RR 05. Cambridge: Oxford and Cambridge Schools Examination Board.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: American Council on Education and Macmillan Publishing Company.

Newbould, C.A. (1982). Subject preferences, sex differences and comparability of standards. *British Educational Research Journal*, *8*, 141–146.

Newbould, C.A., & Schmidt, C.C. (1983). *Comparison of grades in physics with grades in other subjects*. Test Development and Research Unit, RR/83/07. Cambridge: Oxford and Cambridge Schools Examination Board.

Newton, P.E. (1997). Measuring the comparability of standards between subjects: Why our statistical techniques do not make the grade. *British Educational Research Journal*, *23*, 433–449.

Newton, P.E. (2005). Examination standards and the limits of linking. *Assessment in Education*, *12*, 105–123.

Nuttall, D.L., Backhouse, J.K., & Willmott, A.S. (1974). *Comparability of standards between subjects*. Schools Council Examinations Bulletin 29. London: Evans/Methuen.

Partis, M.T. (1997). *Scaling of tertiary entrance marks in Western Australia*. Osbourne Park, Western Australia: Western Australia Curriculum Council.

Pollitt, A. (1996). *The 'difficulty' of A level subjects*. Unpublished research paper, University of Cambridge Local Examinations Syndicate.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Rutter, P. (1994). The effect of studying A level mathematics on performance in A level physics. *Physics Education*, *29*(1), 8–13.

Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.

Seneta, E. (1987). *The University of Sydney scaling system for the New South Wales Higher School Certificate: A manual*. Sydney: Department of Mathematical Statistics, University of Sydney.

Sparkes, B. (2000). Subject comparisons – A Scottish perspective. *Oxford Review of Education*, *26*, 175–189.

UK GCE Examination Boards. (1971). *Dual entry in the 1966 GCE examination. A report prepared by the GCE examination boards in the United Kingdom, January 1971*. Cambridge Assessment Archive, PP/TSW3/8.

Welsh Joint Education Committee. (1973). *Standards in subjects at GCE ordinary level, June 1971*. Research Report No. 1. Cardiff: Welsh Joint Education Committee.

Western Australia Curriculum Council. (1998). *Scaling*. Western Australia: Western Australia Curriculum Council.

Willmott, A. (1995). *A national study of subject grading standards at A level, summer 1993.* Report commissioned by the Standing Research Advisory Committee of the GCE boards.

Wright, B.D., & Stone, M. (1979). *Best test design*. Chicago: MESA Press.

Year 11 Information System. (2006). *Relative ratings*. *Year 11 indicator system*. Durham: Curriculum, Evaluation and Management Centre, University of Durham

Appendix: Equations for the models

General notation used

For simplicity, any subjects taken by only one candidate, or candidates taking only one subject, are excluded from the analysis in all models.

The following symbols and notation have been used across all models.

D_i is the relative difficulty, or grade correction, for subject *i*

A_n is the ability of candidate *n*

D_{ig} is the relative difficulty of grade *g* in subject *i*

F_{ig} is the incremental difficulty of grade *g* relative to the overall difficulty, D_i, of the subject *i*. Hence, $D_{ig} = D_i + F_{ig}$

P_{nig} is the probability that candidate *n* taking subject *i* achieves grade *g*

S_n is the number of subjects taken by candidate *n*

C_i is the number of candidates taking subject *i*

N is the total number of examinations taken in all subjects by all candidates (i.e. $N = \sum_i C_i = \sum_n S_n$)

Φ_i is the set of candidates taking subject *i*

Θ_n is the set of subjects taken by candidate *n*

G_{ni} is the grade achieved by candidate *n* in subject *i* (if taken)

H_{ni} is the mean grade achieved by candidate *n* in their $S_n - 1$ other subjects, so

$$H_{ni} = \frac{1}{S_n - 1}\left(\left[\sum_{j \in \Theta_n} G_{nj}\right] - G_{ni}\right) \quad \text{(provided subject } i \text{ is taken by candidate } n\text{)}$$

μ_i is the mean grade for all candidates in subject *i*, so

$$\mu_i = \frac{1}{C_i} \sum_{n \in \Phi_i} G_{ni}$$

σ_i is the standard deviation of grades for all candidates in subject *i*, so

$$\sigma_i^2 = \frac{1}{C_i} \sum_{n \in \Phi_i} \left(G_{ni} - \mu_i \right)^2$$

Unbiased mean total (UBMT) method

The difficulty, D_i, for each subject, *i*, is defined as the difference between the mean grade achieved in that subject (μ_i) and the mean of all grades achieved in their other subjects by candidates who took subject *i*. Hence,

$$D_i = \mu_i - \frac{\sum_{n \in \Phi_i} \left[H_{ni}(S_n - 1) \right]}{\sum_{n \in \Phi_i} (S_n - 1)}$$

Kelly's method

Grade corrections, D_i, for each subject can be found as the solution to the matrix equation:

$$\boldsymbol{B} \cdot \boldsymbol{d} = \boldsymbol{v}$$

where $\boldsymbol{B}$ is a symmetric matrix $[b_{ij}]$, $\boldsymbol{d}$ is a vector of grade corrections for each subject $[D_i]$, and $\boldsymbol{v} = [v_i]$, such that

$$v_i = \sum_{n \in \Phi_i} (H_{ni} - G_{ni})$$

and

$$b_{ij} = \begin{cases} C_i + \dfrac{C_i^2}{N} & \text{if } i = j \\[2ex] \dfrac{C_i C_j}{N} - \displaystyle\sum_{n \in \Phi_i \cap \Phi_j} \frac{1}{S_n - 1} & \text{if } i \neq j \end{cases}$$

These equations are taken from Lawley's Appendix to Kelly (1976a).

Analysis of variance (ANOVA)

The grade achieved by candidate *n* in subject *i* is modelled as

$$G_{ni} = u + A_n + D_i + e_{ni}$$

where *u* is 'an average grade for all candidates and subjects' (not necessarily the mean as calculated) and e_{ni} is the residual or error term (Nuttall *et al.*, 1974, p. 38).

If the e_{ni}s are further constrained to average 0 for each candidate and for each subject, and the D_is to sum to 0, then all terms are defined uniquely. The solution amounts to the inversion of a *k* x *k* matrix, where *k* is the number of subjects.

Average marks scaling

If G_{ni} is the unscaled (raw) mark awarded to candidate n in subject i, then the (within-subject) standardised mark, Z_{ni}, is given by

$$Z_{ni} = \frac{G_{ni} - \mu_i}{\sigma_i}$$

The scaled mark, X_{ni}, is given by

$$X_{ni} = R_i\, Z_{ni} + D_i$$

where R_i is an index of the relative spread of 'ability' of the students who have taken subject i, and D_i is its relative difficulty.

R_i is defined to be the standard deviation of the Z_{nj}s for all the subjects taken by all candidates who took subject i. Hence,

$$R_i^2 = \frac{\sum_{n\in\Phi_i}\sum_{j\in\Theta_n} Z_{nj}^2}{\sum_{n\in\Phi_i} S_n} - \left(\frac{\sum_{n\in\Phi_i}\sum_{j\in\Theta_n} Z_{nj}}{\sum_{n\in\Phi_i} S_n}\right)^2$$

And the D_is are then calculated by solving the simultaneous equations:

$$D_i \cdot \sum_{n\in\Phi_i} S_n = \sum_{n\in\Phi_i}\sum_{j\in\Theta_n} \left(R_j Z_{nj} + D_j\right)$$

(Derived from Partis, 1997)

Rasch

The partial credit model used here is:

$$\ln\left(P_{nig}/P_{ni(g-1)} \right) = A_n - D_i - F_{ig} = A_n - D_{ig}$$

COMMENTARY ON CHAPTER 9

Iasonas Lamprianou

Heated discussions about the comparability of standards between subjects have worried qualification authorities and the public around the world for many years. Coe's well-thought-out chapter about 'common examinee methods' presents various statistical methods that have been used in the past, and discusses three different conceptions of comparability as applied to examination standards. Coe claims that a comparability method may be validated only under the light of a specific conception of comparability and that 'most of the methods considered [in the chapter] are capable of valid interpretations under the right conditions'. Having this in mind, the author concludes that 'there may be a case, therefore, for the methods that have been outlined in this chapter to play, and be seen to play, a bigger part than they currently do within England in the processes of monitoring and ensuring "comparability"'.

However, international experience has shown that it is often difficult to achieve 'the right conditions' in order to make undisputed interpretations of the results of comparability methods. This commentary briefly reviews three international case studies (based on Lamprianou, 2007) to illustrate that the practical application of comparability methods has been marked with severe and continuous doubts and fierce criticisms by local societies. Contrary to Coe's suggestion for a more systematic use of common examinee methods, this commentary draws a line of caution: more explicit uses of such methods in the English context should probably be done with extra care and only after significant consideration and public debate.

The case of Western Australia

In the case of Western Australia, the Average Marks Scaling (AMS) method (a common examinee method discussed by Coe) is used to accommodate for the fact that students are examined on different subjects in order to be awarded an overall University Admission Index. The Technical Committee on Scaling, a committee responsible for the application of the method, has frequently come forward with admittedly compelling evidence about the fairness of the method but it failed to persuade everybody. The governmental agencies have published widely, albeit in vain, trying to explain the scaling process and to persuade parents and students that there are no problems regarding the comparability between subjects (Universities Admissions Centre, 2006).

The Technical Committee on Scaling (2002) urges 'students… not to try to "work the system" – they are likely to get it wrong' because students often try to identify 'easy' subjects (that are said to be scaled down) in order to avoid them. There is a number of 'frequently asked questions' that come up in governmental leaflets and discussions in the media: 'Are there subjects that are always scaled down?', 'Is it true that if I study this course I can't get a high [scaled score]?' etc. During a recent discussion in the

Parliament (House of Representatives, 2006) it was heard that 'We have… had evidence that if you are studying chemistry you are likely to have your marks downgraded as well' and that 'We have certainly had significant evidence that there is deep concern about the UAI formula'. The scaling method has been accused of threatening the students away from specific subjects and the Technical Committee on Scaling (2002) has commented that 'each year brings its own myths and conspiracy theories'.

The case of Cyprus

The very same 'frequently asked questions' are raised by politicians, parents and students in Cyprus, a country with a long tradition of using scaling methods to make the scores of students on different subjects comparable. Chemistry, a subject notorious among parents and students for allegedly 'consistently being scaled down' compared to other subjects has seen its enrolment dropping by 70% from 2001 to 2006. The source of the problem – as has been repeatedly claimed – mainly lies in the ('qualitatively') different groups of students that actually happen to be examined on the same subjects and in their varying motivations.

It would be very difficult for any comparability method to accommodate for such strong sub-group and motivation effects – except perhaps if different scaling is used for different candidates. At the moment, however, the Cyprus Testing Service has problems persuading the public and politicians about the fairness of the existing comparability method (which one MP has recently characterised as 'complicated formulae… that no one will understand'); how might somebody try to convince them that different methods should be used for different groups of students?

The case of Fiji

The third, and last, example of public distrust comes from the Fiji islands where a statistical method is used to scale the marks of the students in different subjects in order to award a single aggregate score. The Ministry of Education has tried hard to convince the public about the fairness of the method, giving examples of its widespread use abroad; alas, with little success. The minutes from a 2004 session of the Parliament read:

> Currently, there is a strong public debate on the scaling of marks and the general feeling is that, it is a bad practice… The session on the scaling of marks by the Ministry of Education's Exam Office was very revealing… It had a formula that made little sense [and] factors no one knew how they were derived.
>
> Parliament of Fiji (2004)

The Fiji Human Rights Commission was reported in the summer of 2006 to be investigating alleged breaches of human rights when students' external examination marks were scaled. According to Radio New Zealand International (2006), 'the *Fiji Times* reports that this follows complaints that the mark scaling system is unfair, non-transparent and violates the Bill of Rights in the Constitution and the Fiji Human Rights Commission Act'.

A pattern of doubts and distrust

There is a pattern in all three case studies above: (a) the comparability methods in use are difficult for the layperson to understand; (b) there are beliefs of allegedly unfair effects on the system; and (c) there is a general feeling of distrust. But is there a relationship between these issues and the conceptions of comparability suggested by Coe and others in this book?

It is difficult to see how students or parents could argue in favour of the 'performance' conception of comparability: being examined on different subjects, e.g. French or chemistry is obviously a non-comparable experience. If we would ever try to publicly support a performance conception of comparability, we would hit problems; people would not understand our complex web of specifications and assumptions about the intellectual demand made by each of the examinations. This might provoke feelings of unfairness, and we would probably spark a new round of doubt and distrust.

Having said this, however, the stakeholders in the three case studies mentioned above (as well as in other places like Tasmania, Singapore, Canada's British Columbia, etc.) seem to compromise with the idea that there must be something common between all subjects – albeit 'practically' different – so that one is justified in aggregating scores on different subjects for practical ranking purposes. It seems that when the public is faced with the need to make direct comparisons between candidates to allocate scarce educational resources, the 'construct' conception is easier to endorse. The problem, though, with the construct conception of comparability is that it is best (though not solely) served by complex statistical techniques like the Rasch model (used by the Tasmanian Qualifications Authority) or the AMS. Complex methods, however, are incomprehensible to the layperson and even rare examples of 'unfair' scaling results are usually put forward by the media to erode public trust in the system. Such models imply that the success rate will not be the same on different subjects, therefore generating allegations of unfairness (tapping into the concept of 'statistical' comparability).

The statistical conception of comparability initially looks deceptively easier to endorse by parents and students. Intuitively the public will question the results when the pass rates on different subjects are widely different; so would keeping the grade distribution similar on each subject solve the problem of perceived unfairness? Such an approach to comparability could probably involve simpler statistics so the first problem (that the layperson does not understand the statistics) could be solved. However, the media would probably fail such an approach on the grounds of first-page 'case study examples', i.e. bright students being awarded very different grades on different subjects depending on their popularity and the competition within them.

Unravelling the Gordian knot of comparability is very difficult. The public seems to have a dual approach: they generally endorse the construct approach for practical purposes, while they hesitate to drop the statistical approach altogether. Drawing on the experience of other countries, England must be very careful before proceeding with further (and probably formal) use of comparability methods. Since pragmatism should prevail when attempting to solve practical educational problems with social consequences, it is important to seek a general consensus from the major stakeholders before taking the next step.

References

House of Representatives. (2006, March 10). Standing committee on agriculture, fisheries and forestry. Armidale: Commonwealth of Australia, House of Representatives.

Lamprianou, I. (2007). *The international perspective of examination comparability methods. Technical report*.

Parliament of Fiji. (2004, July 29). *Parliamentary debates, House of Representatives*. Daily Hansard.

Radio New Zealand International. (2006, August 31). *Fiji human rights commission to investigate exam scaling*.

Technical Committee on Scaling. (2002). *Report of calculation of universities admission index 2001*. Academic Board Report. New South Wales: Committee of Chais, University of New South Wales.

Universities Admissions Centre. (2006). *You and your UAI. A booklet for 2006 New South Wales HSC students*.

COMMENTARY ON CHAPTER 9

Alastair Pollitt

This commentary concerns the causes of apparent differences in the 'difficulty' of different subjects that appear in common examinee analyses. In Chapter 9 Coe refers several times to Pollitt (1996), but each time only to relatively minor points in that paper, whose main purpose was to report a comparison of the results of common examinee analyses in England and Singapore which helps shed light on some factors that contribute to the pattern obtained. Since that paper has not been published, the gist of the report is repeated here.

Our interest was aroused by discussion surroundings Dearing's (1996) review of 16–19 qualifications, in which the results of common examinee analyses seemed to imply that some A level examinations were consistently harder than others. A very similar pattern had been reported for Scottish exams by Kelly (1976), and since then in other 'old Commonwealth' countries.

The Dearing report recommended that England should 'raise the demand of any subjects found to be decidedly below the average'. But the consistency of the pattern argued against that conclusion. How could it be that Scotland and the other countries somehow took their subject standards from England, where the original standards had been set wrongly? Alternatively, how could every one of these independent assessment systems have accidentally made the same pattern of errors as each other? It seemed to me more reasonable to see the consistent patterns as evidence that England (and the others) had got things more or less right than that everyone had got them wrong. But it was possible that the common cultural foundation of the old Commonwealth could be a consistent source of some kind of bias.

A level results data were available from one other country – Singapore – which had not been analysed in this way before, and which was culturally rather different. In just three decades it had risen from third world status to achieve a GNP per head at least as high as its former colonial master; also, in each of 1995, 1999 and 2003 it was the most successful country in the world in the TIMSS 8th grade mathematics studies. The majority of the population are ethnically Chinese, but with a generally high level of English as a second language, and Europeans are a small minority. Singapore is different from old Commonwealth countries in many ways: would these differences lead to a different pattern?

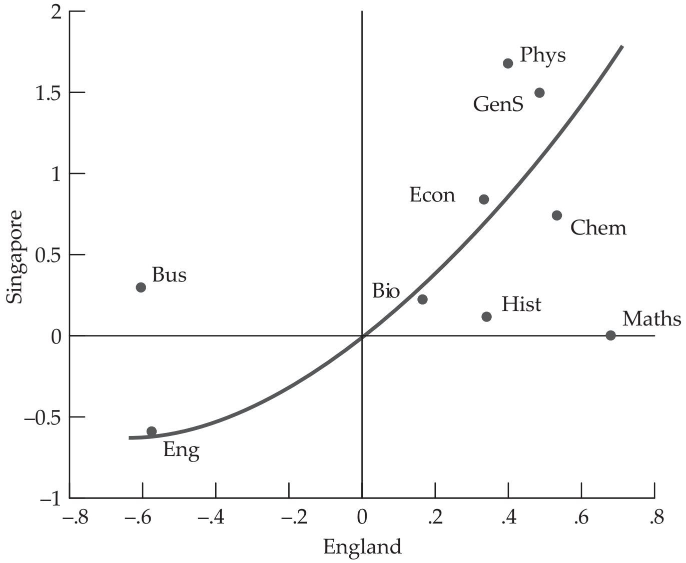
Comparing subject pair analyses

In the early 1990s most of the University of Cambridge Local Examinations Syndicate (UCLES) overseas A level examinations were either identical to the UK ones or constructed in parallel by overlapping teams of examiners. The mathematics exam was identical. The ‘Management of Business’ exam differed from the UK ‘Business Studies’ one in that an extra written paper replaced the course work element, but the common papers were used to set the standard. There is no reason why any of these minor differences should interfere with the analysis here, given that the UK pattern of apparent difficulty holds up so consistently across so many different UK syllabuses.

A subject pairs analysis was carried out using mathematics as a reference. Every other subject for which pair information was available from more than 1,000 candidates per year was compared to mathematics. All analyses are based on the averages of three years, 1993–95.

Figure 1 shows the results of the analysis for Singapore plotted against the ‘corrections’ for England published in Dearing (1996); the scale is grades, so physics seemed to be 1.7 grades ‘harder’ than mathematics for Singaporean students. As in the Dearing and ALIS studies,[1] physics and general studies appear at the top of the order of ‘difficulty’, and English at the bottom.

Figure 1 Relative ‘difficulties’ in Singapore and England



The Singapore differences are clearly larger, the difference between ‘hardest’ and ‘easiest’ (physics and English) being 2.28 grades compared to 1.30 in Dearing. And chemistry has moved substantially, becoming 0.85 grades ‘easier’ than physics (or perhaps physics has become ‘harder’ by the same amount) while in England they seem equally difficult. Nevertheless, with the help of the quadratic regression line plotted on the figure, the Singapore pattern can be interpreted as similar to that observed in other developed Western nations – supporting the view that there is some overall truth in the relative order.

In order to plot this line, however, it is necessary to omit two subjects that stand out as very different, business and mathematics. In the Western countries mathematics is consistently in the 'most difficult' group, but in Singapore it is 'easier' than anything else except English. In fact it was a concern on the part of their Ministry that mathematics might really be too easy that initiated this study. Of course we might consider it remarkable that English should seem so easy, in a country where it is not the native language of the vast majority of candidates. A special educational effort is certainly, and necessarily, put into English language teaching to compensate for this so that school pupils are able to study through the medium of English. It is likely too, though, that the group who choose to take English literature as an A level subject are rather different from their British counterparts.

Why should mathematics seem to be so easy in Singapore? It is common for British newspapers to report that Eastern children are much better than England's at mathematics – with the implication that they are better at everything else as well. The analysis here seems to agree that, at A level at least, they are superior in mathematics – but only in mathematics, since it seems so much more easier relative to the others than in England.

However, another familiar stereotype is that the 'tiger economies' are very business oriented. It is therefore something of a surprise to see that they find 'Management of Business' so much more 'difficult' than Western candidates do. If *they're good at mathematics* explains its relative 'easiness' how does *they're good at business* fit with its relative 'difficulty'? It is of course quite easy to invent a plausible explanation: suppose that 'business oriented' really means that many of them expect to get jobs in business or set up their own businesses, and they therefore take business as an A level even if they are not particularly good at it – perhaps alongside mathematics which they are good at. Perhaps most of the good mathematics students think that they ought to combine mathematics with something practical, so that they will be able to get – or create – proper jobs. The result would be that most mathematics+business students would be better at mathematics than at business.

Such explanations may or may not be true; we have no way of knowing from the statistics alone. The main conclusion of this analysis has to be that we cannot interpret differences between subject mean grades, whether from subject pair analyses or any other analysis, as evidence of differences in 'difficulty' unless we know who took each subject, and why. The 'appeal of the method', as Coe notes, depends heavily on the presumption that perfectly matching the *general* features of the individual candidates is an adequate basis for comparing the two examinations, but the discussion in the last paragraph shows that the *specific* interactions between student and subject – their interest in it, their motivation to succeed, and their reasons for choosing to study it – are likely to affect their level of success in ways that *general* characteristics cannot predict. A similar argument applies at the level of schools: the method presumes that matching samples in terms of the *general* characteristics of the schools they attend is sufficient to equate the students' expected grades. However, when 85% of England's secondary schools are now 'Specialist schools' (DfES, 2007), with *specific* expertise in just one or two out of ten curricular areas, we can no longer assume that any school will provide the same quality of teaching and preparation in every subject.

Demands and conceptions of comparability

The differences between Coe's three kinds of comparability arise from how they consider *difficulty* and *demands* (see Chapter 5 for an extended discussion of how these two concepts can be distinguished). His 'statistical' comparability is wholly concerned with the difficulty (not the demands) of examinations. Evidence of sub-group differences like those reported here, or of sex differences within England, make this conception of comparability very problematic. The use of terms like 'a "typical" student', and the quotation from Nuttall *et al*. (1974), show that it is fundamentally a norm-referenced approach, and inappropriate for an examination context where the samples taking any two examinations will *never* be representative of the same population. Examination boards routinely use common examinee analyses to monitor standards, but they do not act on first-order differences like those discussed here; instead they look for second-order differences – *changes* in the normal pattern – as indicators that something may have gone wrong on a particular occasion.

'Construct' comparability relates wholly to demands (not to difficulty), since the *linking construct* will in fact be the subset of all the demands in either exam that is common to both of them; this common subset may well be unrepresentative of all the demands in either of them. Thus with this conceptualisation it is quite possible, on a subject pairs basis, for A to be 'harder' than B, for B to be 'harder' than C, and for C to be 'harder' than A, since the linking construct will be different for each comparison. If many examinations are included in the comparison the linking construct, being common to all, must be a very small subset of the demands of some of them, which will seriously undermine the usefulness of the comparisons. Further, it is very important to remember that, as in factor analysis, giving a 'label' to the common subset is far from a trivial exercise, and it is very dangerous – to claim that 'here the linking construct is "general aptitudes"' or 'here... might be... the effectiveness of the teaching received' unless you have compelling empirical evidence to support the label.

Coe's 'performance' comparability involves a mixture of difficulty and demands. The phrase 'would be seen as more "difficult"... if it made a greater demand on the candidate' shows that examiners who judge comparability with this conception must consider both the demands the examination made and how hard the candidates found it to meet them. The two examinations will not contain the same questions, and the judges must estimate how much 'ability' is needed to produce the observed quality of performance after compensating for their estimates of the levels of demands in the tasks. 'Ability' in this sense is estimated on some scale that the judges have formulated intuitively as the *common* scale for the two (or more) exams. As with 'construct' comparability, this common-scale comparison is then used to infer the standard on the rest of each examination even though the rest of their demands may well be quite different. This is certainly the most complex of the three conceptions: it requires the judges to combine different kinds of evidence in a sophisticated way that cannot easily be modelled statistically or even described theoretically.[2] The real world of examining is complex in just this way, and we should not be surprised that 'construct' comparability 'appears to be the default in thinking about comparability'.

In passing, we explain in Chapter 5 why the view that 'it should be possible to specify the difficulty of an examination even if no candidate has ever taken it' will for ever remain an illusion.

To summarise, we cannot get rid of the sub-group anomalies just by switching to a different conceptualisation of comparability; they are real and problematic. The demands of A level mathematics are the same whether you are a boy from Singapore or a girl from England but, because of a different cultural setting and the expectations and exam preparations that result from it, the difficulty of the exam is quite different for the two groups.

Endnotes

[1] The ALIS studies were reported in Dearing (1996).

[2] Indeed, Good and Cresswell (1988) provided evidence that even the wisest examiners may be systematically biased in making these judgements.

References

Dearing, R. (1996). *Review of qualifications for 16–19 year olds*. London: School Curriculum and Assessment Authority.

Department for Education and Skills. (2007). *The standards site. Specialist schools: What are specialist schools?* Available at: http://www.standards.dfes.gov.uk/specialistschools/

Good, F.J., & Cresswell, M.J. (1988). Grade awarding judgments in differentiated examinations. *British Educational Research Journal*, *14*, 263–281.

Kelly, A. (1976). A study of the comparability of external examinations in different subjects. *Research in Education*, *16*, 37–63.

Nuttall, D.L., Backhouse, J.K., & Willmott, A.S. (1974). *Comparability of standards between subjects*. Schools Council Examinations Bulletin 29. London: Evans/Methuen.

Pollitt, A. (1996). *The 'difficulty' of A level subjects*. Unpublished research paper, University of Cambridge Local Examinations Syndicate.